

## Laboratorio de Datos - Guía de ejercicios Calidad de Datos



## Objetivo.

Ejercitar los conceptos de Calidad de Datos.

## Ejercicio 1

He aquí algunas situaciones reales en que la mala calidad de los datos trajo pérdidas económicas, algunas de ellas fácilmente cuantificables (una vez producidas).

**Caso 1.** Pozo petrolero perforado en una ubicación errónea por interpretación equivocada del sistema de coordenadas en uso. La empresa fue multada.

**Caso 2.** Un banco local fue condenado a pagar a un cliente indemnizaciones por cientos de miles de pesos por haber sido incluido erróneamente en bases de datos de morosos.

Fuente: Diario Clarín 14/02/2003

https://www.clarin.com/economia/banco-debe-pagar-120000-incluir-mal-cliente-veraz 0 rJ4iuMIAFq.html

**Caso 3.** En un organismo del gobierno de un país latinoamericano se mandaron cartas a todas aquellas empresas beneficiadas por una norma. El 30% de la correspondencia volvió rechazada por problemas en la dirección.

- a) Para cada uno de los casos:
  - Identificar quiénes fueron los afectados en cada situación (usuarios o clientes, managers que hacen uso de los datos, desarrolladores o encargados de mantenimiento de los sistemas, otros)
  - ii. ¿Qué impacto identifica en estos casos (además del económico descrito)?
    (descreimiento en la organización, causa de costos innecesarios, impacto en toma de decisiones, disminución de satisfacción de usuarios y clientes, etc.)
- **b)** Describir algún inconveniente de Calidad de Datos que lo haya afectado en su vida personal y/o alguno que haya detectado a nivel laboral.

# Ejercicio 2

Dar al menos dos ejemplos de sistemas o conjuntos de sistemas con pocos bugs, pero que permitan el almacenamiento de información con problemas de calidad.

## Ejercicio 3

Dados los siguientes inconvenientes clasifíquelos según el origen de los mismos (instancia, proceso, modelo, software):

- a. Datos obligatorios que no se asumen como tales y por lo tanto no se cargan
- b. Interfaces poco amigables
- c. Rangos de valores que no se respetan
- d. Distintas personas cargan la misma información haciendo distintas asunciones
- e. Gente que hace modificaciones pero no debería estar autorizada para hacerlas
- f. Hay información que no está presente porque no hay forma de almacenarla



# Laboratorio de Datos - Guía de ejercicios Calidad de Datos



- g. El mundo que se quiere representar evolucionó, pero esta situación no se ve reflejada en el sistema
- h. Datos que han cambiado en el mundo real, y que no fueron actualizados
- i. Datos que provienen de distintas fuentes y que no son consistentes
- j. Datos correspondiente al año, que han sido almacenados con dos dígitos en lugar de cuatro
- k. Posibles valores completados en el campo región:
  - ANETOFAGASTA
  - ANMTOFAGASTA
  - ANTOFAGASTA
  - ANTO9FAGASTA
  - ANTOAFAGASTA
  - ANTOFAAGASTA

### Ejercicio 4

Dados los siguientes problemas, i) clasificarlos en función del atributo de calidad que se ve afectado; ii) determinar si el problema es de modelo o de datos.

- a. No se cargan unidades de medida en que se midió la profundidad de un pozo petrolero
- b. No es posible almacenar el sistema de referencia
- c. Hay inconsistencias entre nombres de un mismo pozo petrolero en distintos sistemas
- d. La ubicación de una central telefónica no coincide con la ubicación real
- e. El nombre de un pozo petrolero no corresponde con el que debería ser, de acuerdo a la ley
- f. Hay personas fallecidas que figuran como empleados participantes de cursos (por los cuáles la empresa que los informa consigue una exención impositiva)
- g. Las direcciones de los clientes no están actualizadas

En los casos del **Ejercicio 3** originados por instancia o modelo, mencione qué atributos de calidad se ven afectados.

#### Ejercicio 5

A modo de repaso de temas vistos anteriormente

- a. ¿Qué problemas de diseño encuentra en la tabla que figura a continuación?
- b. ¿Qué tipo de anomalías produce?
- c. ¿Qué problemas de calidad de datos puede acarrear?
- d. ¿Qué otros problemas de diseño (de la base de datos) cree que pueden afectar la calidad de los datos? Relacionarlos con los atributos de calidad del modelo
- e. ¿En los casos en que se deje información redundante en una tabla adrede, cómo recomienda proceder para evitar problemas en la calidad de los datos?

**Nota:** RUT es el Rol Único Tributario (el número con el cuál se identifica a las personas físicas y jurídicas en Chile, similar a nuestro CUIT/CUIL).



## Laboratorio de Datos - Guía de ejercicios Calidad de Datos



CUOTAS A VENCER

RUTEMPRESA: VARCHAR2 (42)

RAZON: VARCHAR2(210) RUTTRAB: VARCHAR2(42) NOMTRAB: VARCHAR2(101) NOMCOMUNA: VARCHAR2(64)

CODCOMUNCA: NUMBER VALORCUOTA: NUMBER

# A continuación se muestra un posible conjunto de datos de esta tabla

				•		•			
RUTEMPRESA	RAZON				RUTTRAB	NOMTRAB	NOMCOMUNA	CODCOMUNA	VALORCUOTA
2178645-4	Servando	Humberto	Arriagada	Peres	10734185-4	LUIS ALFREDO CASTILLO	TEMUCO	93801	32000
2178645-4	Servando	Humberto	Arriagada	Perez	12192576-1	Cesar Enrique Castillo	TEMUCO	93802	32000

### Ejercicio 6

Un cliente desea conocer la calidad de sus datos en cuanto a:

- a. Empresas sin dirección que posee almacenadas en su sistema
- b. Empresas que parecerían estar almacenadas más de una vez en su sistema (puede asumir que se identifican por nombre más dirección)
- 1. Definir métricas según el modelo Goal Question Metric (GQM) para dar soporte al cliente. Recuerde que en todos los casos debe identificar el objetivo, la pregunta y la métrica.
- 2. ¿En algún caso le puede ser de utilidad el uso de algún algoritmo de matching para ejecutar la métrica?
- 3. Determinar en qué casos es conveniente el uso de algoritmos específicos (por ej. soundex y keyboard distance) para la ejecución de las métricas.

#### Ejercicio 7

Tomar el dataset corregido de Dengue DatosDengueYZikaCorregida.csv (del campus virtual) y listar los nombres de departamento y sus ids, nombres provincia e id provincia para todos aquellos departamentos con mismo nombre, pero distinto id de departamento y distinto id de provincia. Ordenarlos por nombre de departamento.